International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Classification of Diabetes Using Cuckoo Search Optimization in Big Data Environment

N.V.Poornima[1]*, Dr.B.Srinivasan*[2]*, Dr.P.Prabhusundhar*[3]

[1,2,3] Department of computer science Gobi Arts & Science college Gobichettipalayam,India

ABSTRACT

Diabetes can be turn into life threaten diseases, if it is not treated at an early stage. Especially, in the women, the chance of diabetes is higher as compared to men due to the hormonal changes during pregnancies. Due to this, they suffer long term diabetes as well as other diseases due to their tensions and regular life chores. This can be prevented if the diagnosis is determined at an early stage. Mostly, the trained doctors are required to confirm the diabetes. It requires manual work and complete knowledge in it. This problem is avoided by several research works using machine learning algorithm for the classification. Many algorithms are there to process effectively for smaller dataset with smaller number of attributes. But for the huge amount of data still question mark (?). Hence, to overcome this short coming, an optimization based classifier is proposed called cuckoo search optimization; the objective function is to reduce the misclassification rate of the classifier. This approach able to improve the accuracy as compared to the existing technique.

Keyword: Diabetic, optimization, cuckoo search.

Introduction to Big Data

The term Big in “Big Data” itself defines it operations and size is large. Big indicates not only the size, it also denotes the nature and processing of the data. Big data doesn’t mean it refers to only large amount of data. But, the size of data can be small but it is complex in nature and requires tedious process to perform an operation.

It is a special type of data whose volume is in exponentially increasing manner with respect to time. It requires special type of processor to store, retrieve and perform operations on it. Moreover, it cannot be processed with the simple data mining algorithms. It requires special tool and algorithms for processing it. All operations can be performed on the big data as similar to the normal data processing methods.

Big data can also be called as a combination of three V’s. The three V’s are Variety, Velocity and Volume. These three V’s are change in nature and it will be of higher volume for the big data.

Here, in variety, there are three types. They are structured, unstructured and semi-structured. The term structured indicates that data is arranged in an ordered format. The unstructured data denotes there is no definite format for the data. In Semi-structured, it is a hybrid of both structured and unstructured format.

Big data plays a major role in many applications like healthcare units, Educational oriented, Banking sector, Information technology sector, Manufacturing and retail sector. In health care unit, it plays a greater role by storing the information in a larger and continuous manner for saving the patient information. Not only for the storing purpose, it able to help the doctors by diagnosis the diseases by combining big data and artificial intelligence algorithms. In this, the role of big data in diagnosis the diabetics using machine learning technique. The remaining section describes about the techniques used in the big data for predicting the diabetes.

Introduction to cuckoo search

Increasingly more current metaheuristic calculations motivated essentially are arising and they become progressively well known. For instance, particles swarm improvement (PSO) was propelled by fish and bird swarm insight, while the Firefly Algorithm was enlivened by the blazing example of tropical fireflies. These nature-motivated metaheuristic calculations have been utilized in a wide scope of streamlining issues, including NP-difficult issues like the mobile sales rep issue. The force of practically all cutting edge metaheuristics comes from the way that they impersonate the best component in nature, particularly organic frameworks developed from regular determination more than a long period of time. Two significant attributes are choice of the fittest and variation to the climate. Mathematically talking, these can be converted into two significant attributes of the advanced metaheuristics: escalation and broadening. Escalation means to look around the flow best arrangements and select the best applicants or arrangements, while broadening ensures the calculation can investigate the hunt space productively. This paper expects to detail another calculation, called Cuckoo Search (CS), in view of the intriguing reproducing bebaviour, for example, brood parasitism of specific types of cuckoos. We will initially present the reproducing bebaviour of cuckoos and the attributes of Levy trips of certain birds and natural product flies, and afterward form the new CS, trailed by its execution. At last, we will contrast the proposed search technique and other well-known advancement calculations and talk about our discoveries and their suggestions for different enhancement issues.

Cuckoo Search (CS) is an evolutionary optimization algorithm which is inspired by Yang and Deb(2009) . The theory of cuckoo search was inspired by the species of bird called cuckoo.Cuckoo is a fascinating bird, not only because of the beautiful sounds they can make, but also becauseof their aggressive reproduction strategy, where a mature cuckoolaid their eggs in other host birds orspecies. This is known as obligate brood parasitis.The basic of this algorithm is thespecific egg laying and breeding of cuckoos itself. In this case if a host bird discovers the eggs are notits own, it will either throw these alien eggs away or simply abandon its nest and build a new nestelsewhere.
For simplicity in describing the Cuckoo Search, three idealized important rules are used:

i) Each cuckoo lays one egg at a time, and dumps its egg in a randomly chosen nest.
ii) The best nest with high quality of eggs will carry over to the next generation.The number of available hosts’ nests is fixed
iii) The egg laid by a cuckoo is discovered by the hostbird with a probability *pa* ε (0, 1).

The Pseudo code for the CSO is given below.

```
start
objective function
initialization parameters like host nest, iterations boundary and number of cuckoo
while (iterations< max(iterations))
   select randomly the cuckoo and its solution using Levy flight equation 6
   evaluate objective function
   choose nest randomly
if (current fitness > previous fitness)
   Replace the previous solution with current solution
End if
Remove worse nest based on $P_a$ and find new solution using equation 6
  Find local best solutions
  Rank them and the top best solution is the global best solution.
End while
Output= dominant attribute and corresponding
Fitness function value
End
```

For a maximization problem, the quality or fitness of a solution can simply be proportional to the value of the objective function. Other forms of fitness can be defined in a similar way to the fitness function in genetic algorithms. For simplicity, we can use the following simple representations that each egg in a nest represents a solution, and a cuckoo egg represent a new solution, the aim is to use the new and potentially better solutions (cuckoos) to replace a not-sogood solution in the nests. Of course, this algorithm can be extended to the more complicated case where each nest has multiple eggs representing a set of solutions. For this present work, we will use the simplest approach where each nest has only a single egg. Based on these three rules, the basic steps of the Cuckoo Search (CS) can be summarized as the pseudo code shown in Fig. 1. When generating new solutions x (t+1) for, say, acuckoo i, a Levy flight is performed

$$x\ (t+1)\ i = x\ (t)\ i + \alpha \oplus \text{Levy}(\lambda) \text{-----}\ (1)$$

Where $\alpha > 0$ is the step size which should be related to the scales of the problem of interests. In most cases, we can use $\alpha = 1$. The above equation is essentially the stochastic equation for random walk. In general, a random walk is a Markov chain whose next status/location only depends on the current location (the first term in the above equation) and the transition probability (the second term). The product $\oplus$ means entrywise multiplications. This entrywise product is similar to those used in PSO, but here the random walk via Levy flight is more efficient in exploring the search space as its step length is much longer in the long run. The Levy flight essentially provides a random walk while the random step length is drawn from a Levy distribution Levy~ $u = t^{-\lambda}$, $(1 < \lambda \leq 3)$, (2) which has an infinite variance with an infinite mean. Here the steps essentially form a random walk process with a power-law step-length distribution with a heavy tail. Some of the new solutions should be generated by Levy walk around the best solution obtained so far, this will speed up the local search. However, a substantial fraction of the new solutions should be generated by far field randomization and whose locations should be far enough from the current best solution, this will make sure the system will not be trapped in a local optimum. From a quick look, it seems that there is some similarity between CS and hill-climbing in combination with some large scale randomization. But there are some significant differences. Firstly, CS is a population-based algorithm, in a way similar to GA and PSO, but it uses some sort of elitism and/or selection similar to that used in harmony search. Secondly, the randomization is more efficient as the step length is heavy-tailed, and any large step is possible. Thirdly, the number of parameters to be tuned is less than GA and PSO, and thus it is potentially more generic to adapt to a wider class of optimization problems. In addition, each nest can represent a set of solutions; CS can thus be extended to the type of meta-population algorithm
The new solution for a cuckoo is determined with the help of levy flight. Levy flight is used to provide random walk for the cuckoo and it follows power law step length distribution. Based on the above Pseudo code, the optimization will determine the best attribute for the classification and it will be used for the training and testing.

Highlights on Cuckoo Search

CS is a new evolutionary optimization algorithm which is inspired by lifestyle of bird family. This section presents about CS based on the statistical result from the papers that were reviewed. According to the statistical results CS algorithm has been applied in different kinds of optimization problems across various categories. the major categories considered in Cuckoo Search Algorithm were Engineering followed Pattern Recognition, Software Testing & Data Generation, Networking, Job Scheduling and Data Fusion and Wireless Sensor Networks.

Attribute selection using Cuckoo search optimization

This section is to describe the process of selecting the dominant attribute in the outlier removed dataset. The dominant attribute selection is performed to reduce the computational time for the calculation.

The optimization process is to find the solution for a problem through iteration process. The cuckoo search optimization is used. Because, the cuckoo lays it egg in other birds stronger nest and hatch its egg. Due to this, the selection of nest is chosen by satisfying all the conditions. In this, the condition is to maximize the accuracy of the classifier. The objective function of the cuckoo search optimization is given in equation

$$OF_{CSO} = 1 - Accuracy$$

The new solution for a cuckoo is determined with the help of levy flight and it is given in equation 6.

$$X_i^{t+1} = X_i^t + \alpha \oplus Levy\ (\lambda), \alpha\ is\ step\ size, \alpha > 0,$$

$$\mathrm{L}evy = t^{-\lambda}, 0 < \lambda < 3$$

Levy flight is used to provide random walk for the cuckoo and it follows power law step length distribution.

Based on the above Pseudo code, the optimization will determine the best attribute for the classification and it will be used for the training and testing.

Conclusion

Diabetes can be turn into life threaten diseases, if it is not treated at an early stage. Especially, in the women, the chance of diabetes is higher as compared to men due to the hormonal changes during pregnancies. Due to this, they suffer a long term diabetes as well as other diseases due to their tensions and regular life chores. This can be prevented if the diagnosis is determined at an early stage. Mostly, the trained doctors are required to confirm the diabetes. It requires manual work and complete knowledge in it. This problem is avoided by several research works using machine learning algorithm for the classification. Those algorithms process effectively for smaller dataset with smaller number of attributes. Hence, to overcome this shortcomings, in this an optimization based classifier is proposed to process on larger datasets like BIG DATA. The proposed optimization helps to select an optimal attribute which is enough to predict diabetic or non-diabetic using feed forward neural network classifier.

REFERENCES

[1]Barthelemy P., Bertolotti J., Wiersma D. S., A L´evy flight for light, Nature, 453, 495-498 (2008).

[2] Bonabeau E., Dorigo M., Theraulaz G., Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, (1999)

[3] Blum C. and Roli A., Metaheuristics in combinatorial optimization: Overview and concepturalcomparision, ACM Comput. Surv., 35, 268-308 (2003).

[4] Brown C., Liebovitch L. S., Glendon R., L´evy flights in DobeJu/'hoansi foraging patterns, Human Ecol., 35, 129-138 (2007).

[5] Chattopadhyay R., A study of test functions for optimization algorithms, J. Opt. Theory Appl., 8, 231-236 (1971).

[6] Deb. K., Optimisation for Engineering Design, Prentice-Hall, New Delhi, (1995).

[7] Gazi K., and Passino K. M., Stability analysis of social foraging swarms, IEEE Trans. Sys. Man. Cyber. Part B - Cybernetics, 34, 539-557 (2004).

[8] Goldberg D. E., Genetic Algorithms in Search, Optimisation and Machine Learning, Reading, Mass.: Addison Wesley (1989).

[9] Kennedy J. and Eberhart R. C.: Particle swarm optimization. Proc. of IEEE International Conference on Neural Networks, Piscataway, NJ.pp. 1942-1948 (1995).

[10] Kennedy J., Eberhart R., Shi Y.: Swarm intelligence, Academic Press, (2001).

[11] Passino K. M., Biomimicrt of Bacterial Foraging for Distributed Optimization, University Press, Princeton, New Jersey (2001).

[12] Payne R. B., Sorenson M. D., and Klitz K., The Cuckoos, Oxford University Press, (2005). [13] Pavlyukevich I., L´evy flights, non-local search and simulated annealing, J. Computational Physics, 226, 1830-1844 (2007).

[14] Pavlyukevich I., Cooling down L´evy flights, J. Phys. A:Math. Theor., 40, 12299-12313 (2007).

[15] Reynolds A. M. and Frye M. A., Free-flight odor tracking in Drosophila is consistent with an optimal intermittent scale-free search, PLoS One, 2, e354 (2007).

[16] Schoen F., A wide class of test functions for global optimization, J. Global Optimization, 3, 133-137, (1993).

[17] Shang Y. W., Qiu Y. H., A note on the extended rosenrbock function, Evolutionary Computation, 14, 119-126 (2006).

18] Shilane D., Martikainen J., Dudoit S., Ovaska S. J., A general framework for statistical performance comparison of evolutionary computation algorithms, Information Sciences: an Int. Journal, 178, 2870-2879 (2008).